Appling Parallelism in Image Mining

J. Fernández, N. Miranda, R. Guerrero, F. Piccoli
Computational Inteligence Research Laboratory
University of San Luis
Ejército de los Andes 950, San Luis, Argentina
{*jmfer, ncmiran, rag, mpiccoli*}@*unsl.edu.ar*

March 26, 2007

Abstract

Image mining deals with the study and development of new technologies that allow accomplishing this subject. A common mistake about image mining is identifying its scopes and limitations. Clearly it is different from computer vision and image processing areas. Image mining deals with the extraction of image patterns from a large collection of images, whereas the focus of computer vision and image processing is in understanding and/or extracting specific features from a single image. On the other hand it might be thought that it is much related to content-based retrieval area, since both deals with large image collections. Nevertheless, image mining goes beyond the simple fact of recovering relevant images, the goal is the discovery of image patterns that are significant in a given collection of images. As a result, an image mining systems implies lots of tasks to be done in a regular time. Images provide a natural source of parallelism; so the use of parallelism in every or some mining tasks might be a good option to reduce the cost and overhead of the whole image mining process.

At this work we will try to draw the image minnig problem: its computational cost, and to propose a possible global or local parallel solution.

Keyword: Image mining, image indexing and retrieval, object recognition, image clustering, association rule mining, parallel systems, parallel techniques.

1 Image Mining

The tremendous growing of computerized information volume and variety has triggered the development of new data processing tools, World Wide Web technology and databases technologies that enable inferring useful knowledge from an important data bulk. As a result, it is necessary to support big collections of complex type information which includes complex objects data, spatial information or multimedia information. Many research works have focused on images and image mining.

The most general misinterpretation is that image mining only involves applying already existing data mining algorithms on images. Investigations in the area are usually pointed out into two main directions. The first one involves specific authority applications focusing on extracting most relevant image features, so they could be used in data mining [8][11][12]. The second direction applies to general applications, where the aim is discovering image patterns that might be useful in the understanding of existing interactions between human perception of the images at high level and image features at low level. Investigations in this direction try developments with major certainty of success in recovered images from a general purpose database[7][13][17].

Human visual system has the ability to extract significant image relations which are not represented in low-level primitive image features. Complex information and its use on specific applications leads to describe new association rules to information. The big challenge in image mining is the extraction of implicit knowledge, image data relationship, or other features not explicity stored in a pixel representation. As knowledge representation method, *patterns* have already been used by human being for simulating diverse cognitive process like intuition, intention and thinking. As long as the use of patterns can make the cognitive process more effective, they can be included to describe the complexity and features of objects. Image databases containing raw image data as information, cannot be directly used for image mining purposes. Relational databases, traditionally used in data mining, do not satisfy this need; that is why other types of databases are defined like spatial, temporary, documentary and multimedia databases [18].

Since in image mining the aim is to generate all significant patterns without any knowledge of the image content, the patterns types are diverse. They could be classification patterns, description patterns, correlation patterns, temporal patterns and spatial patterns. Image mining deals with all aspects of large image databases including indexing schemes, image storages, and image retrieval, all concerning in an image mining system[10].

The development of an image mining system is often a complex process since it implies joining different techniques ranging from image retrieval and indexing schemes up to data mining and pattern recognition. Besides, it is expected that a good image mining system provides users with an effective access into the image repository at the same time it recognizes data patterns and generates knowledge underneath image representation. Such system basically should assemble the following functions: image storage, image processing, feature extraction, image indexing and retrieval and, pattern and knowledge discovery.

Figure 1 shows a general structure model for image mining System. The system considers a specified sample of images as an input, whose image features are extracted to represent concisely the image content. Besides the relevance of this mining task, it is essential to consider invariance problem to some geometric transformations and robustness with respect to noise and other distortions in designing a feature extraction operator. After representing the image content, the model description of a given image - the correct semantic image interpretation - is obtained. Mining results are obtained after matching the model description with its complementary symbolic description. The symbolic description might be just a feature or a set of features, a verbal description or phrase in order to identify a particular semantic.

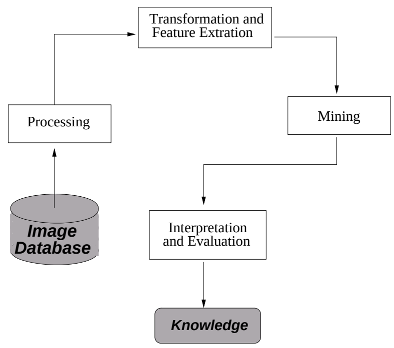

Figure 1: General Image Mining System

2 System Relevant Aspects

Image mining system has two main themes. The first is mining large collections of images and the second is the combined data mining of large collections of images and associated alphanumeric data. The following subsections will try to sketch the involved concepts.

2.1 Minable Images

To get information from a huge amount of images is a hard task. A vast collection of image data should be mined to discover new and valuable knowledge. The next sections explain some relevant techniques applied to convert raw data into minable one.

2.1.1 Image Content Descriptor

Raw images consist of a two dimensional array of pixels, usually named *iconic* format. An image database containing raw information can not be used for mining purposes. All low level information must be preprocessed and transformed into a new suitable format. Numerous image mining techniques are based - explicitly or implicitly - in image models redefined at different representation levels. These new representations try to reflect invisible image features established by association, resemblance, or convention; what is called *symbolic* representation.

Given the variety of known image formats and the significant quantity of information every image represents - explicitly o implicitly -, it is necessary to select the only format representing the required information for knowledge processing. Different alternatives have been proposed [3][4][9], all of them based on converting every image into a feature vector. Problem arises in determining relevant information that must be stored in the vector. The following are the most common criteria used at current developments:

1. *One feature*: only one image characteristic is considered to image arrangement ignoring any another features. It is a simple and efficient alternative. The features most widely used are *color*, *shape* and *texture*.

2. *Multiple features*: a combination of two or more features is taken into account. These can be the own image features - color, shapes, textures, edges and temporary details - , the specific application - medicine, satellite, forensic science - or both. Some researchs combine coarse features like *color*, *texture* and *edge*, while others consider image specific features as *average color*, *standard color deviation*, *distortion* and *grey level closer to the average*, and specific application features like *cardiothoracic radius*, *right and left costofrenic angle*.

3. *The use of transformations*: they allow to express the whole image information in terms of a function. Some transforms are *Discrete Wavelet*, *Gabor Wavelet*, *Multivariated Gaussian model*, among others. These techniques express the image in different frequency components. They have many advantages as linear time, reduction space complexity, reduction of the redundant information providing representantions that make the mining process efficient and accurate .

Since human perception is a subjective sensation, it could be necessary to use multiple representations and feature combination representing all image features and its relationships at low level and high level. At this new representation, it is necessary to reflect the semantic observed in images, which is not possible to detect through a raw analysis of pixel values [15].

2.1.2 Segmentation and Clustering

During preprocessing stage, an image set is selected from image database to be the minable view. The most difficult task of this process is to know the image domain. Sometimes, it is necessary a human expert who makes the required image adjustment and selection or just a query is executed on the database to retrieve the training images. Most of the time, once images are selected, segmentation and clustering techniques are applied to divide images into significant blocks for mining. These techniques have different criteria:

- Image segmentation is a process which partitions, in an exhaustive way, an input image into regions or objects of interest. Each region is considered to be homogeneous in respect to some important feature. Segmentation totally depends on the scene to be sensed, the imaging geometry, the sensors used for the conversion into a digital image and the desired output.

- Clustering algorithms for automatic learning have been applied for image segmentation and computer vision. Data clustering consists on splitting input data into similar objects groups. All the members of a group or cluster are homogeneous among themselves considering some particular feature (color, texture or shape), while different clusters are unlike between them. Image clustering is done usually during the first step of mining process. There exists a spectrum of clustering techniques: *hierarchical clustering*, *partition based methods*, *evolutionary methods*, *mixture resolving*, *nearest neighbor clustering*, among others[1].

  Clustering advantages includes a better image storage and management, an effective indexing method for a quick and efficient retrieval. Interesting characteristics for image mining systems.

Several applications have been developed including segmentation and/or clustering methods during input data preparation stage.

2.2 Combining Tasks: Recovery and Association

All existing mining paradigms depend on input and output data types [16]. Input data types consider two kinds of image representation: iconic o symbolic. Any adopted paradigm considers as input image set a certain image class representing specific semantic entities: objects, concepts, among others and the possible relationships defined among the entities and its attributes.

1. *Content Based Recovery*: it finds in the databases those images similar or semantically equivalent to an input image. The aim is to resolve the recovery by means of keyword-based methods that consider human perception features, so users must express their queries in a semantic way. This kind of task have two disadvantages: low level of retrieval accuracy and high response time for huge dimensional data spaces, resulting in poor performance systems.

2. *Pattern Recognition*: it involves the use of a database with reference images representing diverse semantic entities and its different visions. As a result of a query the user obtains the entity or requested object or a "no" answer if the database does not contain it. Patterns also can be used to represent complex knowledge combining diverse levels of abstraction. This is a process that allows new pattern discovery or to validate the existing ones.

3. *Image model description*: consist of transforming the iconic description of an image into the symbolic one using a particular model structure and a set of parameters.

Besides input and output data types, two other basic tasks must be kept in account: knowledge search and extraction. To develop the first one, the majority of content base image recoveries are applied. For the second one, an input sample image is specified and a symbolic description of the sample image is get as the output [6][14].

3 Proposal

Image mining is a relatively new, wide and very promising field of investigation. Image mining process implies different steps, most of them demanding so many resources and computational time. Parallel processing has become an importante topic when the object is to increase the computational speed of a task [5]. As tasks found their processing in images, and images provide a natural source of parallelism, resolving a global or partial image mining process thru parallel techniques seems to be a good choice.

Many issues of image mining can be optimized with different parallel techniques. Furthermore depending on tasks properties, different parallel paradigms could be applied in the same system. At a first glance, parallel applicant tasks will be: image storage, image processing, feature extraction, image indexing and retrieval and, pattern and knowledge discovery.

As a starting-point, a parallel library with a number of SIMD and MIMD algorithms performing common image processing tasks is been developed. Among of them are: image smoothing, histogramming, 2-D FFT calculation, local area histogram equalization, local area, brightness and gain control, feature extraction, maximum likelihood classification, contextual statistical classification, image correlation (convolution, filtering), scene segmentation, clustering feature enhancement, rendering, etc.[2]

Depending on problem specifications, if the parallel solution follows parallel systems basic principles, the future is very promising.

References

[1] P. Berkhin. Survey of clustering data mining techniques. Technical report, Accrue Software, San Jose, CA, 2002.

[2] A. Choudhary and S. Ranka. Parallel processing for computer vision and image understanding. *IEEE Computer*, 25(2):7–9, 1992.

[3] P. Foschi, D. Kolippakkam, H. Liu, and A. Mandvikar. Feature extraction for image mining. In *Workshop on Multimedia Information Systems (MIS 2002)*, Tempe, Arizona, October 2002.

[4] K. Fukuda and P.A. Pearson. Data mining and image segmentation approaches for classifying defoliation in aerial forest imagery. In *Environmental Modelling and Software (for Sustainability) - iEMSs*, pages 50–58, Burlington, Vermont, USA, July 2006.

[5] A. Grama, A. Gupta, G. Karypis, and V. Kumar. *Introduction to Parallel Computing*. Addison- Wesley, 2003.

[6] Q. Iqbal and J. K. Aggarwal. Lower-level and higher-level approaches to content-based image retrieval. In *IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI 2000)*, pages 197 – 201, Austin, Texas, April 2000. IEEE, IEEE.

[7] Y. Keiji. Managing images: Generic image classification using visual knowledge on the web. In *Proceedings of the eleventh ACM international conference on Multimedia*, pages 167–176, November 2003.

[8] R. Kosala and H. Blockeel. Web mining research: a survey. *ACM SIGKDD Explorations Newsletter*, 2(1):1–15, June 2000.

[9] T. Li, Q. Li, S. Zhu, and M. Ogihara. A survey on wavelet applications in data mining. *SIGKDD Explor. Newsl.*, 4(2):49–68, 2002.

[10] R. Missaoui and R. M. Palenichka. Effective image and video mining: an overview of model-based approaches. In *MDM '05: Proceedings of the 6th international workshop on Multimedia data mining*, pages 43–52, New York, NY, USA, 2005. ACM Press.

[11] T. Mitchell, R. Hutchinson, M. Just, R.S. Niculescu, F. Pereira, and X. Wang. Classifying instantaneous cognitive states from fmri data. In *Proc. 2003 American Medical Informatics Association Annual Symposium*, pages 465–469, 2003.

[12] H. Orallo, R.Quintana, and F.Ramirez. *Introduccion a la Mineria de Datos*. Prentice Hall, 2004.

[13] A. Selim, K. Krzysztof, T. Carsten, and M. Giovanni. Interactive training of advanced classifiers for mining remote sensing image archives. In *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 773–782, 2004. (Industry/government track posters).

[14] L. Silva, O. R. P. Bellon, R. Lemes, J. A. Meira, and M. N. L. Cat. An image processing tool to support gestational age determination. In *CBMS '06: Proceedings of the 19th IEEE Symposium on Computer-Based Medical Systems*, pages 867–874, Washington, DC, USA, 2006. IEEE Computer Society.

[15] P. Stanchev. Using image mining for image retrieval. In *Computer Science and Technology*, may 2003.

[16] A. Vailaya, A.T. Figueiredo, A.K. Jain, and H.J. Zhang. Image classification for content-based indexing. *IEEE Transactions on Image Processing*, 10(1):117–130, Jan 2001.

[17] Y. Wang, F. Makedon, J. Ford, L. Shen, and D. Goldin. Image and video digital libraries: Generating fuzzy semantic metadata describing spatial relations from images using the r-histogram. In *Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*, pages 202–211, 2004.

[18] J. Zhang, W. Hsu, and M. Li Lee. Image mining: Trends and developments. *Journal of Intelligent Information Systems*, 19(1):7–23, 2002.